

# LHCb: Online Compute system for Run 3

Niko Neufeld  
[niko.neufeld@cern.ch](mailto:niko.neufeld@cern.ch)

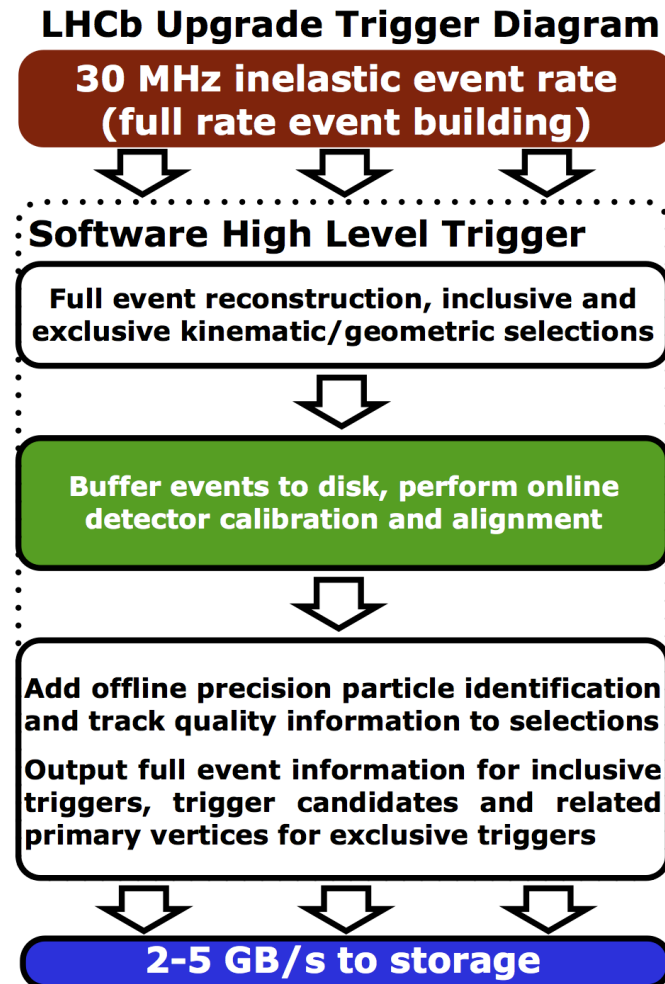
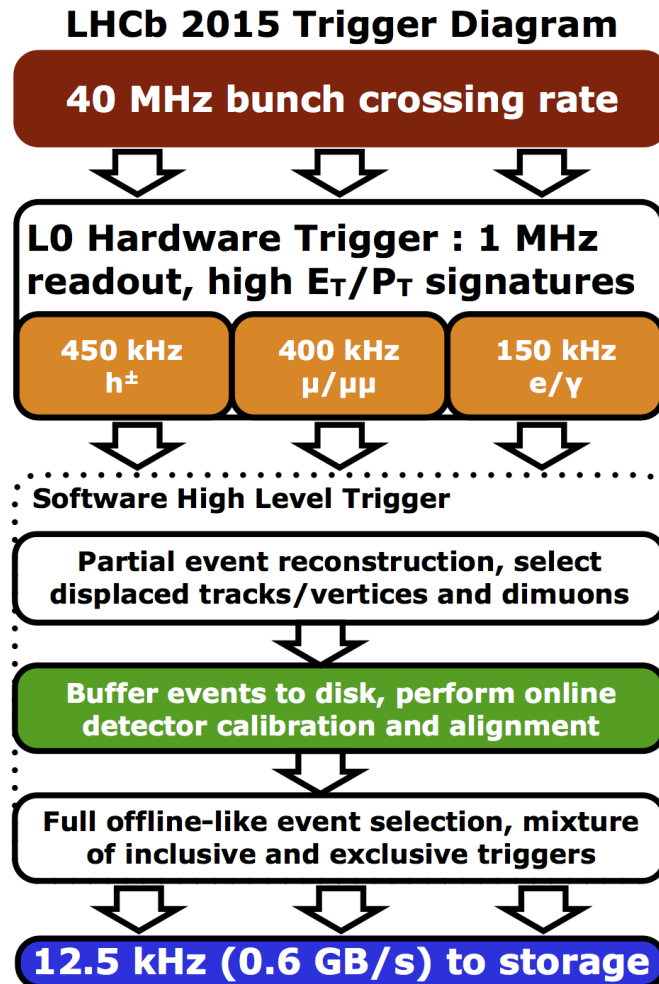


# Outline



- Readout system
  - Slow & fast control
  - Optical links
  - Readout board
  - Event building
  - Data centre

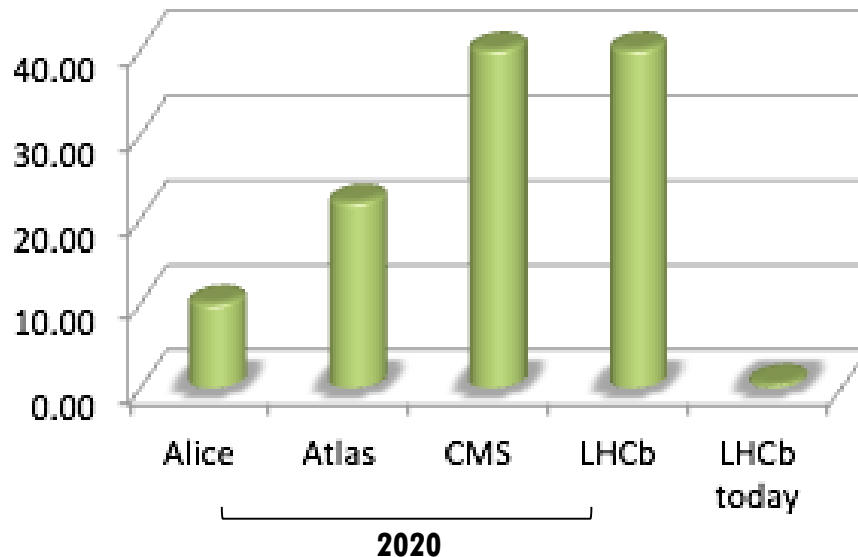
# LHCb Trigger in from Run2 to Run3



# Run3 upgrade

- Filter farm will need to handle:
  - Event size (~130 kB) (@ 30 MHz)
  - Larger event rate (40MHz == LHC bunch crossings per second)
- New challenges for DAQ & High-Level Trigger

Network – Projected Throughput [Tbit/s]



# Run3 Online System

## Dimensioning the system:

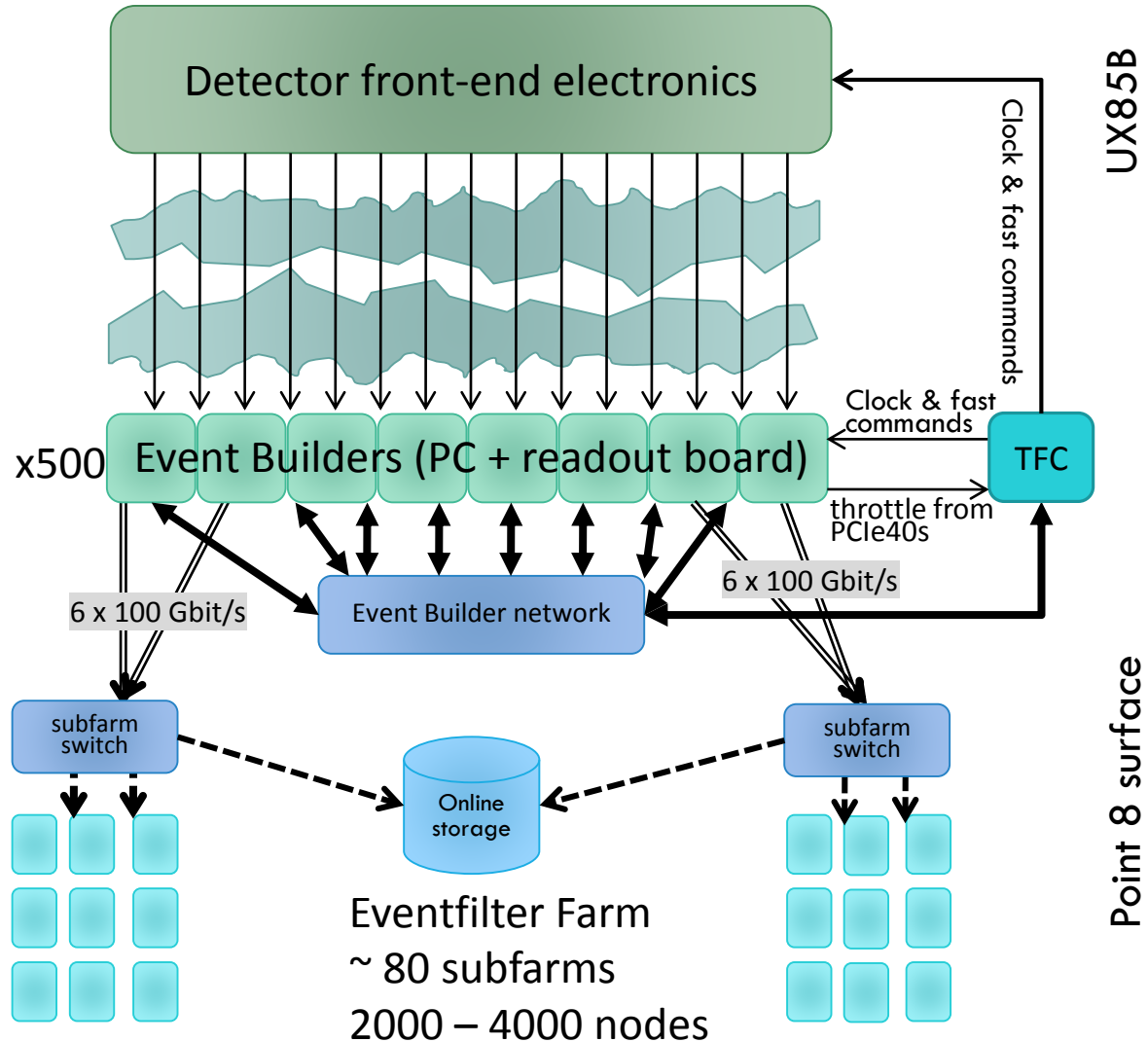
- ~10 000 versatile links
  - detector to surface (~350 m)
  - up to 4.8 Gbps/link
- ~500 readout nodes (up to ~24 links/node)
- ~40 MHz event rate
- ~130 kB event size (@ 27 MHz)

## High bisection bandwidth in event builder network

- ~32 Tb/s aggregate bandwidth
- Leverage 100G Lan technologies

## Global configuration and control via ECS subsystem

## Global synchronization via TFC subsystem



# Numbers

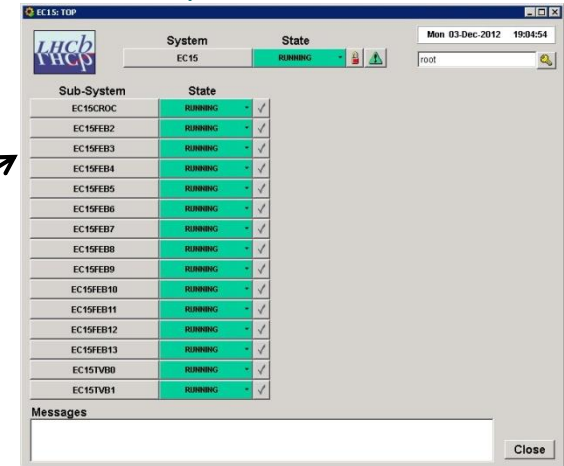


- Event-size (@  $2 \times 10^{33}$ )  $\sim$  130 kB
- Readout-rate 40 MHz
- 500 event-builder nodes
- Between 1000 and 4000 event-filter nodes
  - Dual-socket, accelerator to be decided
- 500 port minimum event-building network
  - TDB: OPA, IB, Ethernet
- 1500 – 4500 port filter network
  - Ethernet?
- New data-centre
  - 4000 rack-units
  - 2 MW max
- 50 to 100 real nodes for “slow” and “fast” control
  - Using PCIe40 cards
- Rest of control-system on virtual machines as today
- Local storage on each filter-unit at least 20 TB  $\rightarrow$  will depend on disk-technology
- Central buffer storage  $\sim$  1 to 2 PB
- $\sim$  10000 uni-directional fibres for DAQ (4.8 Gbit/s)
- $\sim$  2000 fibre-pairs for ECS/TFC (GBT)

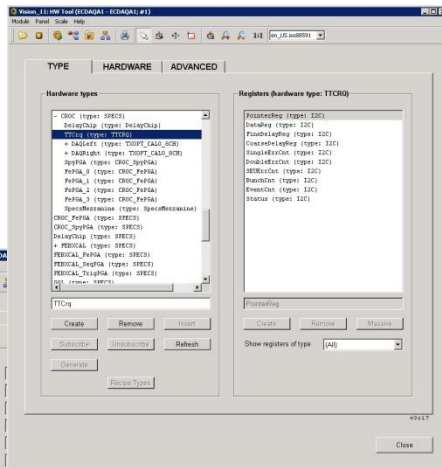
# Experiment Control System for Run 1, 2, 3, 4, 5, 6 ...

- Controls and monitors all subsystems
  - DAQ, TFC, HLT, farm...
- Continuity from current implementation
  - JCOP / DIM / WinCCOA / SMI++ / Recipes
- Already able to drive current readout board prototype, from input to output
- Frontends rely on GBT-SCA hardware by EP-ESE
- Low-level components are being implemented

## Operation UI



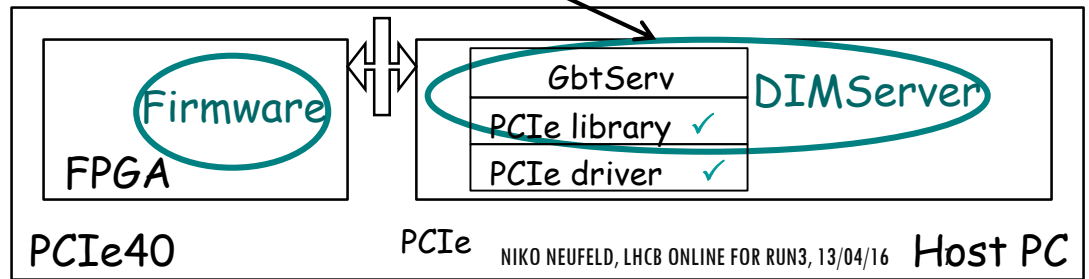
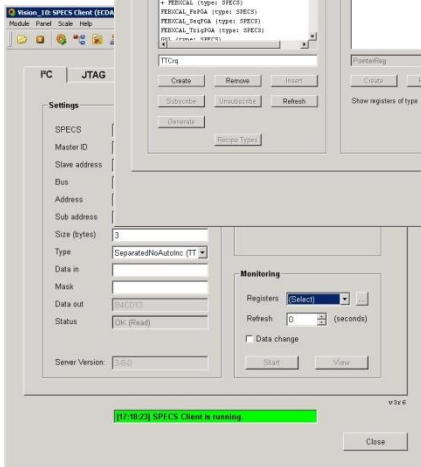
## HW Description UI



## Control PC

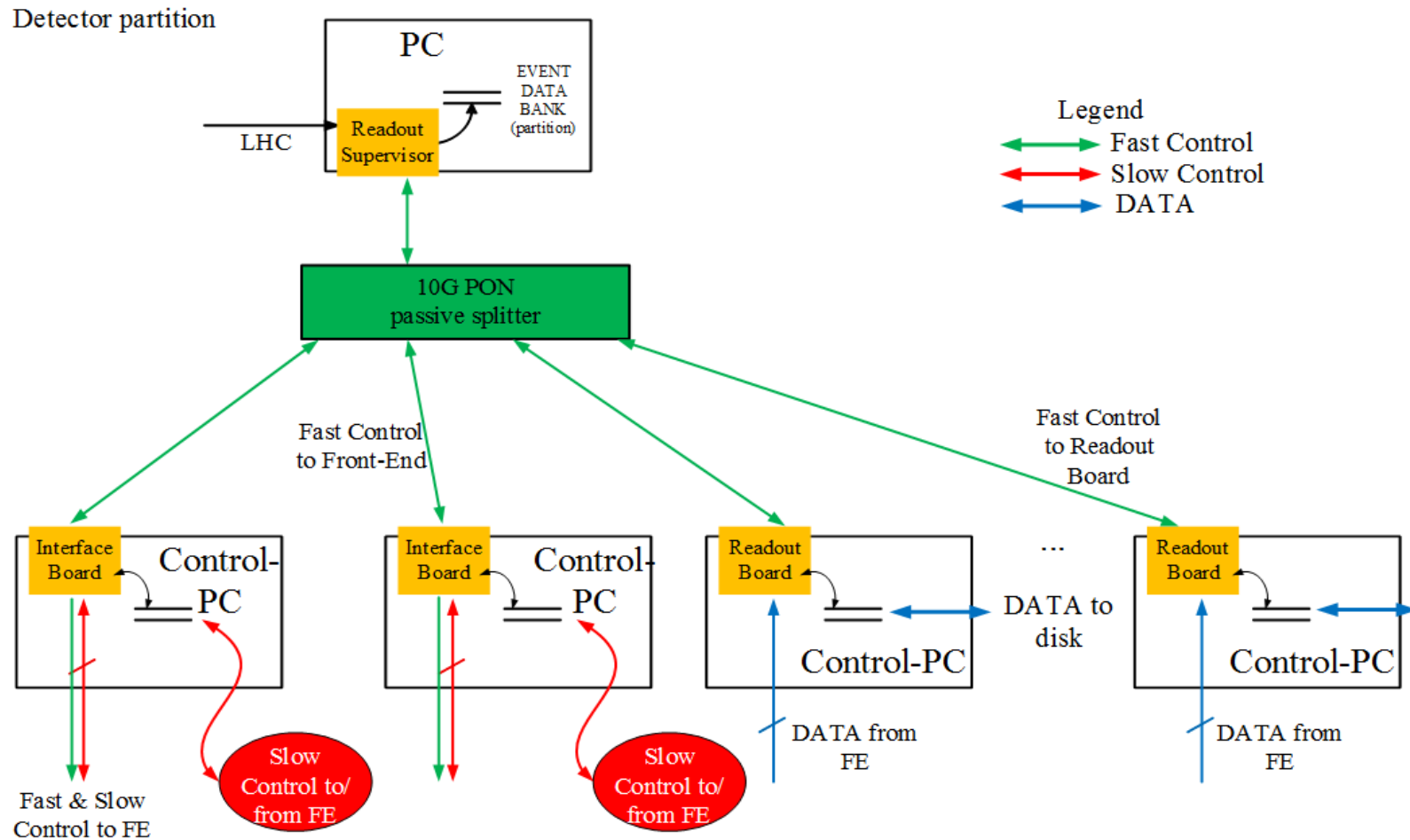


## Test UI





# TFC architecture





# TFC (Timing & Fast Control)

## Current status

- Already integrated in firmware
- Uses same readout board hardware as the DAQ (PCIe40)
- Can send fast commands to frontends
  - SciFi, UT, Muon ASICs already being tested
- Programmable internal throttle for bandwidth regulation

## Ongoing work

- PON (Passive Optical Network) technology integration (with EP-ESE)
- Clock phase tests on readout board (with CPPM)
- Continue feedback and compliance testing with frontend experts

**Commands (TFC\_DEV - TFC\_DEV #1)**

DeviceName: SOL40\_GBTestLink0 | Version: 1.12 | Date: 20150811.02 | State: RUNNING GBT CONNECTED

**Command SM**

- SOL40 -> SODIN Offset: 0
- SOL40 -> TELL40 Offset: 0
- SOL40 -> FE Offset: 3400
- TFC -> SOL40 Delay: 0

**Counters**

- BXD Reset: 41636425
- EID Reset: 0
- FE Reset: 0
- BE Reset: 0
- Header Only Cmd: 0
- NZS Mode Cmd: 0
- BX VETO Cmd: 3185568972
- Snapshot Cmd: 2921
- Sync Cmd: 0
- MEP Accept Cmd: 258199326
- Triggers received: 3872988392

**Register Table (SODIN Single Shots)**

Register	Address	Write	Read
mcCr18	22	0	0
mcCr19	23	0	0
mcCr2	5	0	0
mcCr20	24	0	0
mcCr21	25	0	15
mcCr22	26	0	0
mcCr23	269	0	88
mcCr24	270	0	0
mcCr25	271	0	0
mcCr26	272	0	0
mcCr3	7	0	0
mcCr4	8	0	0
mcCr5	9	0	0
mcCr6	10	0	0
mcCr7	11	0	0
mcCr8	12	0	0
mcCr9	13	0	88

**Commands (TFC\_DEV - TFC\_DEV #1)**

DeviceName: SODIN\_GBTest.Cone0 | Version: 1.06 | Date: 20131204.03 | State: RUNNING

**Statistics and status**

Orbits	Periodic Trig A	Periodic Trig B	5.92 kHz
231741192	115321728	115321727	5.92 kHz
0x000	115321727	115321725	5.92 kHz
994249443	115321725	0.00 kHz	0.00 kHz
4274240377	0.00 kHz	0.00 kHz	0.00 kHz
4274240313	0.00 kHz	0.00 kHz	0.00 kHz
110334 kHz	0	0	0
last Trig Loss	0.077 %	0	0
Sync Cmd	10	0	0
Snapshot Cmd	10181	0	0
BX VETO Cmd	3461820091	0	0
Header Only Cmd	5907	0	0
NZS Mode Cmd	0	0	0
FE Reset	3	0	0
BE Reset	0	0	0
TFC Reset	2	0	0

**Triggers**

Orbit	Raw	Raw Rates (kHz)	Gated	Gated Rates (kHz)
231741192	11846	11.846	0	0
0x00000000	0	0	0	0
Random A	3330184876	1094.893	1316511453	1090.552
Random BB	2752005312	7862.651	0	0
Random B1	1703414102	87.231	0	0
Random B2	1703404487	87.275	0	0
Random EE	3498862891	179.872	0	0
Random C	62496274	31.972	432626733	31.957
Random D	59563392	3.063	41355685	3.058
Periodic 1	115321728	5.923	80177287	5.923
Periodic 2	115321727	5.923	80177288	5.923
Calibration Trig A	115321725	5.923	80177288	5.923
Calibration B	0	0	0	0
Calibration C	0	0	0	0
Calibration D	0	0	0	0
Auxiliary	4274240477	1103.338	4274240313	1695.007

**TFC Functions**

- Periodic Trig 1:
- Periodic Trig 2:
- Calibration Trig A:
- Calibration Trig B:
- Calibration Trig C:
- Calibration Trig D:
- Random Generator:
- Random Trig A:
- Random Trig B:
- Random Trig C:
- Random Trig D:
- NZS Mode:
- NZS Consecutive:
- Snapshot:
- Sync:
- BX Veto:
- Header Only:
- Lumi Trig:
- Event ID Ctrl:
- MEP Destination:
- Dynamic MEP Dest:
- BX Type:

**Calibration Triggers**

Calib A BXID	Calib A periodicity	Calib B BXID	Calib B periodicity	Calib C BXID	Calib C periodicity	Calib D BXID	Calib D periodicity
3087	3087	2	2	2527	2527	2	2
2	2	2	2	2	2	2	2
1199	1199	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2

**Offsets**

Physics Trigger	3450	3459
Auxiliary Trigger	512	512
NZS/TAE latency	13	13
OUT latency	4	4

**Trigger Types**

Periodics	Calibrations	Randoms	NZS	Luminosity	Physics	Beamgas	Auxiliary
9	9	4	4	7	2	1	14
10	10	4	4	2	1	1	14
7	7	2	2	0	0	0	0
2	2	0	0	1	1	1	14
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	14
14	14	14	14	14	14	14	14

**Orbit clock**

External/Internal Orbit	Ext	Int
Orbit length	3564	3564
NZS/TAE latency	100	100
External orbit missing	0	0
Orbit desynchronization	0	0
Orbit presence	0	0

# DAQ cost optimisation

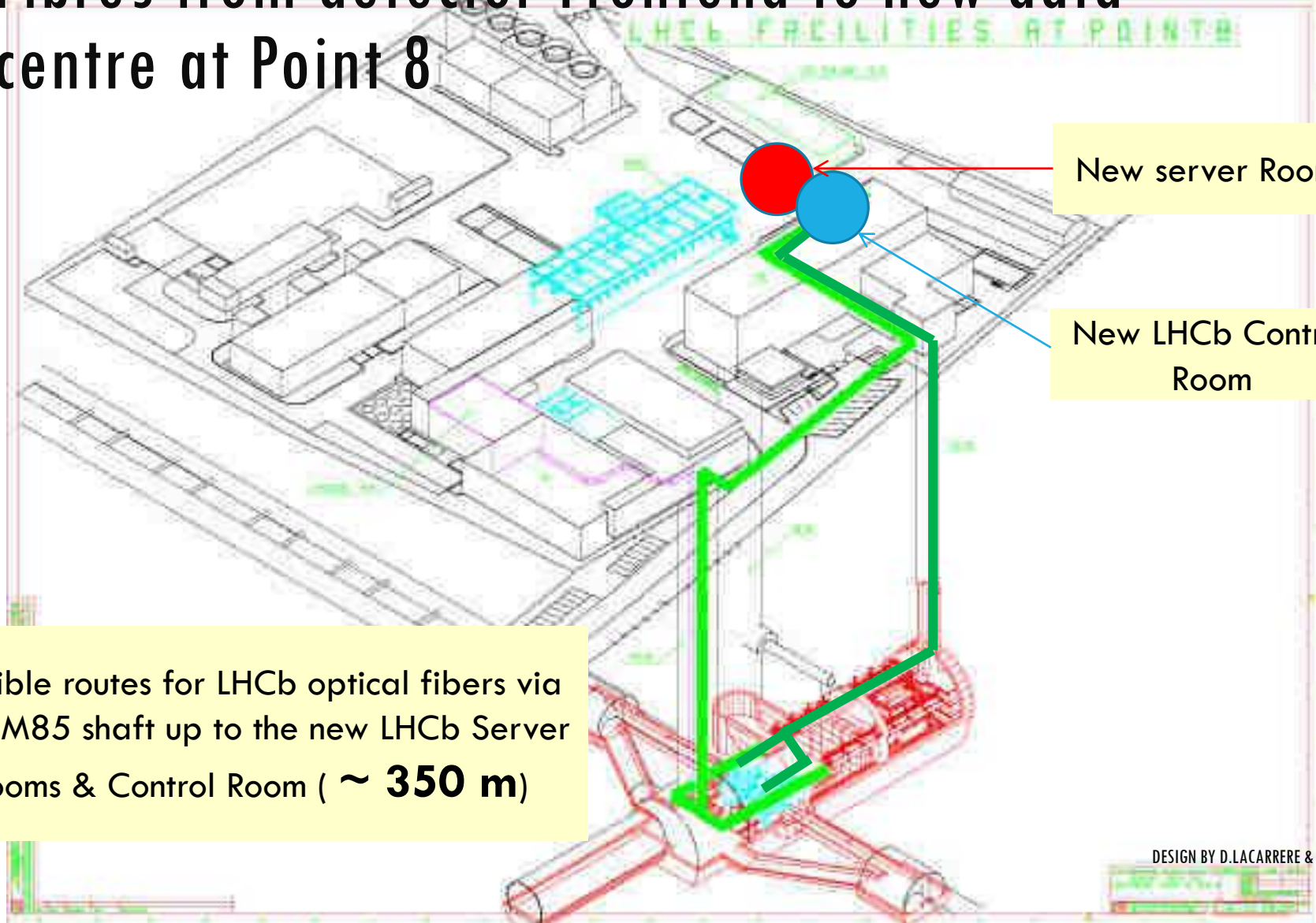


Main cost driver (apart from CPU) is the number and length of fast interconnects (like in HPC!)

Versatile links are there and optical anyhow

Most compact system has everything: event-builder, controls, TFC and farm in one place → new data-centre at the surface and run versatile links to the surface

# Fibres from detector Frontend to new data-centre at Point 8



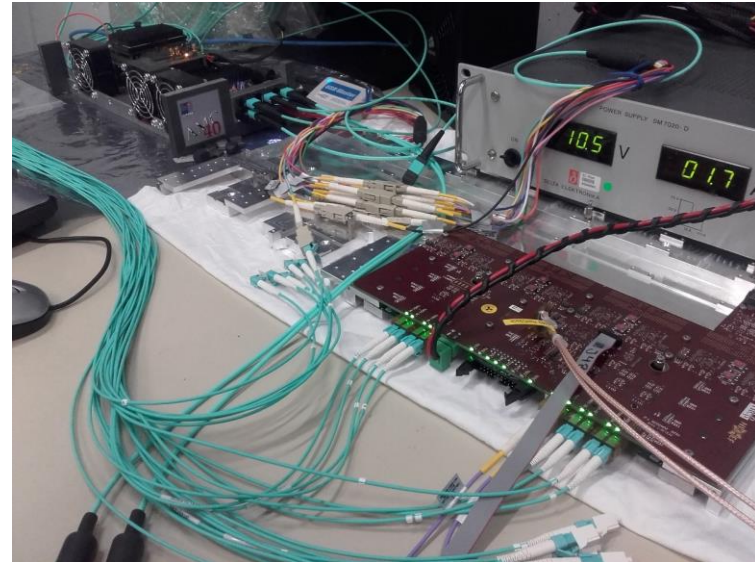
DESIGN BY D.LACARRERE & L.ROY



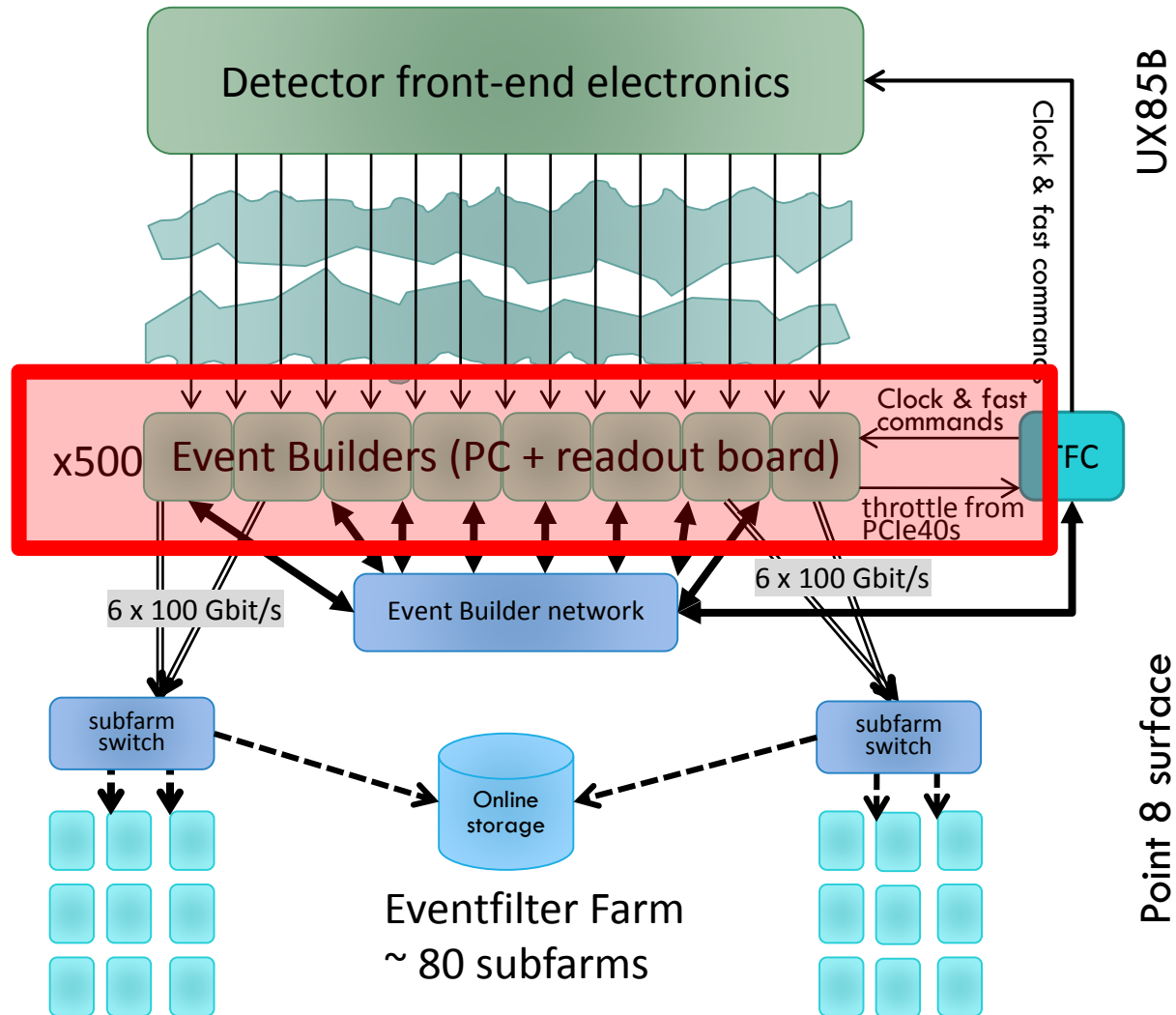
# Long-distance optics



- Counting room on surface
  - Power, cooling, space constraints in underground area
  - ~350 meter distance over OM3 MMF
- Based on EP-ESE technology
  - Rad-hard Versatile Link on frontends
  - Initially qualified for ~100m
- Fiber infrastructure by EN-EL
  - Pilot installation at end 2014
- Loopback tests in 2015
  - ~9 months, ~700 meters
  - Avago MiniPOD transceivers
  - Bit Error Rate  $< 10^{-18}$
  - Full system equivalent:  $< 5$  errors/day
  - ✓ LHCC milestone
- Continued tests in 2016
  - Versatile TX on frontend prototype
  - MiniPOD RX on readout board prototype

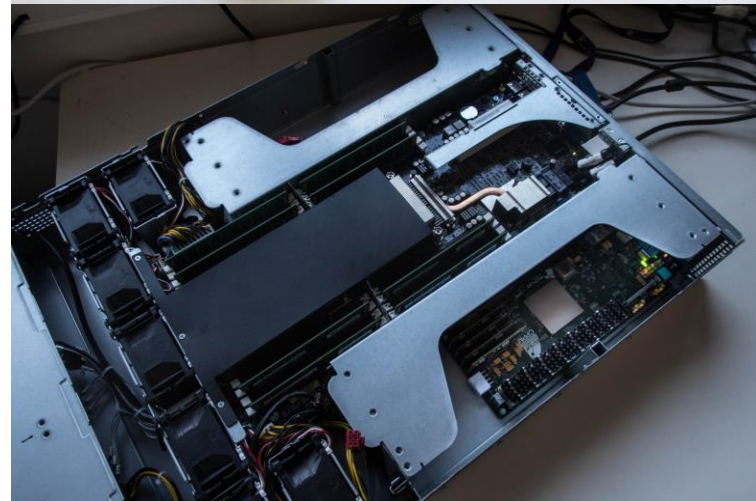
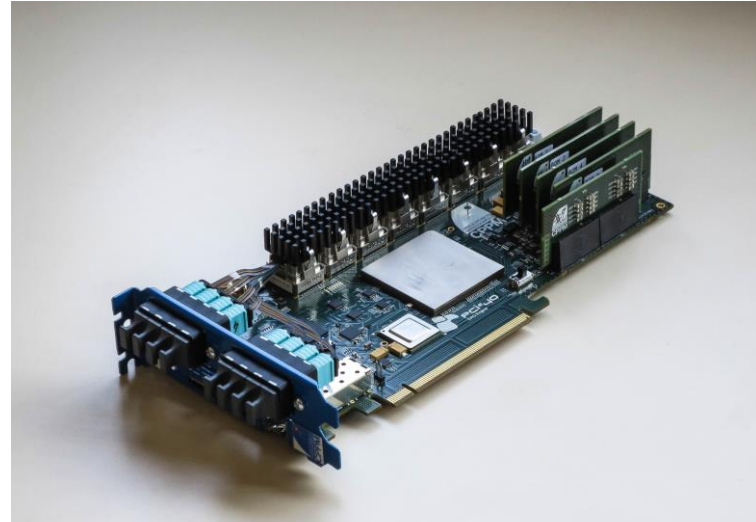


# Readout boards / Event builders

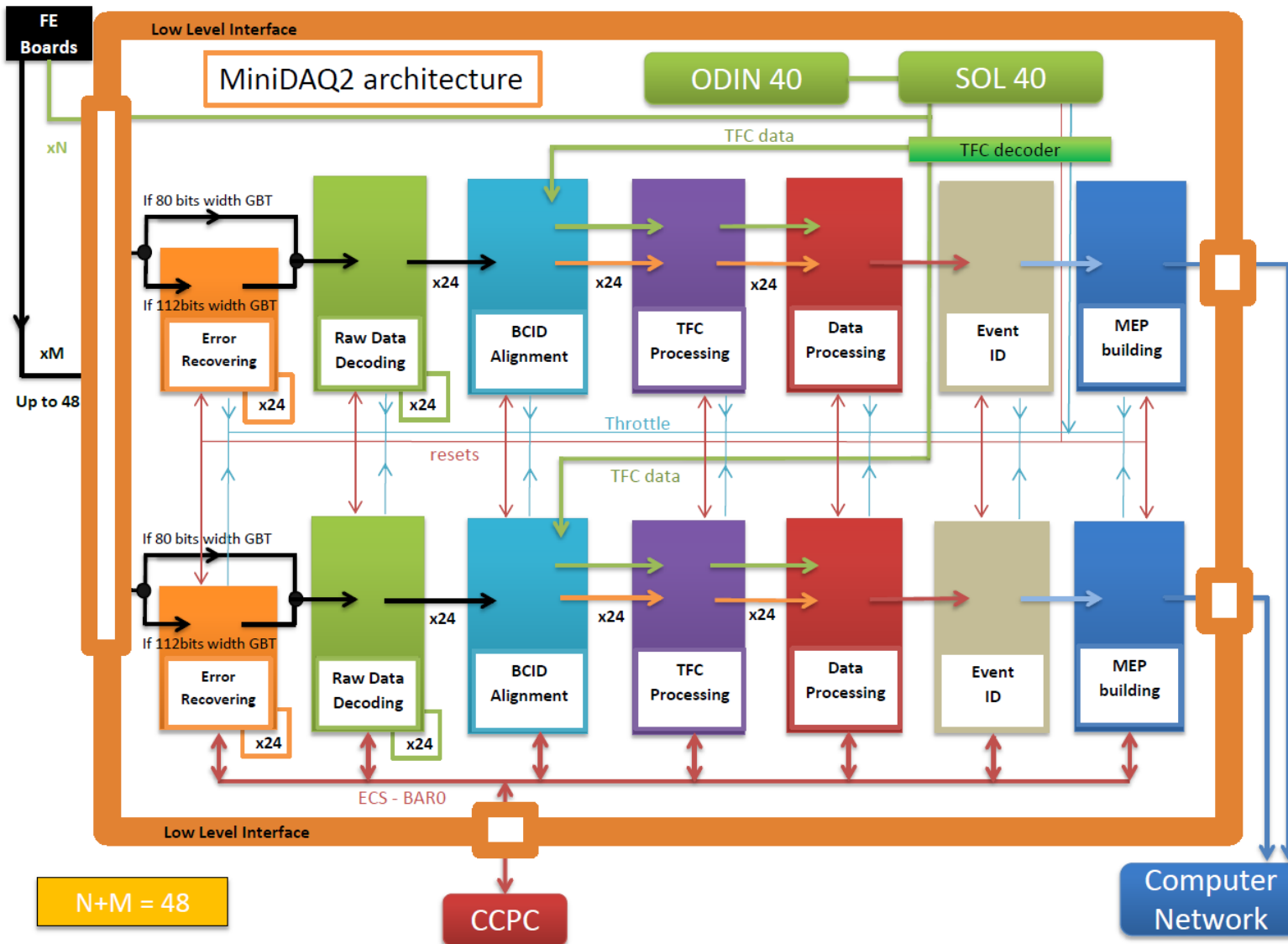


# Readout board hardware (PCIe40)

- PCI Express add-in card
  - Altera Arria10 FPGA
  - 100 Gbps DMA engine to event-builder memory
- High-density optical IO
  - Up to 48 transceivers (Avago MiniPODs)
  - Reuse same HW for timing distribution system
- Decouple FPGA from network
  - Maximum flexibility in network technology
- Exploit commercial technologies
  - PCI Express Gen3 interconnect
  - COTS servers designed for GPU acceleration
- 2<sup>nd</sup> generation readout board
  - Developed at CPP Marseille
- Pre-production launched
  - Ready end of Q2
- Market survey completed
  - Tender in H2 2016
- More in the talk by P. Durante



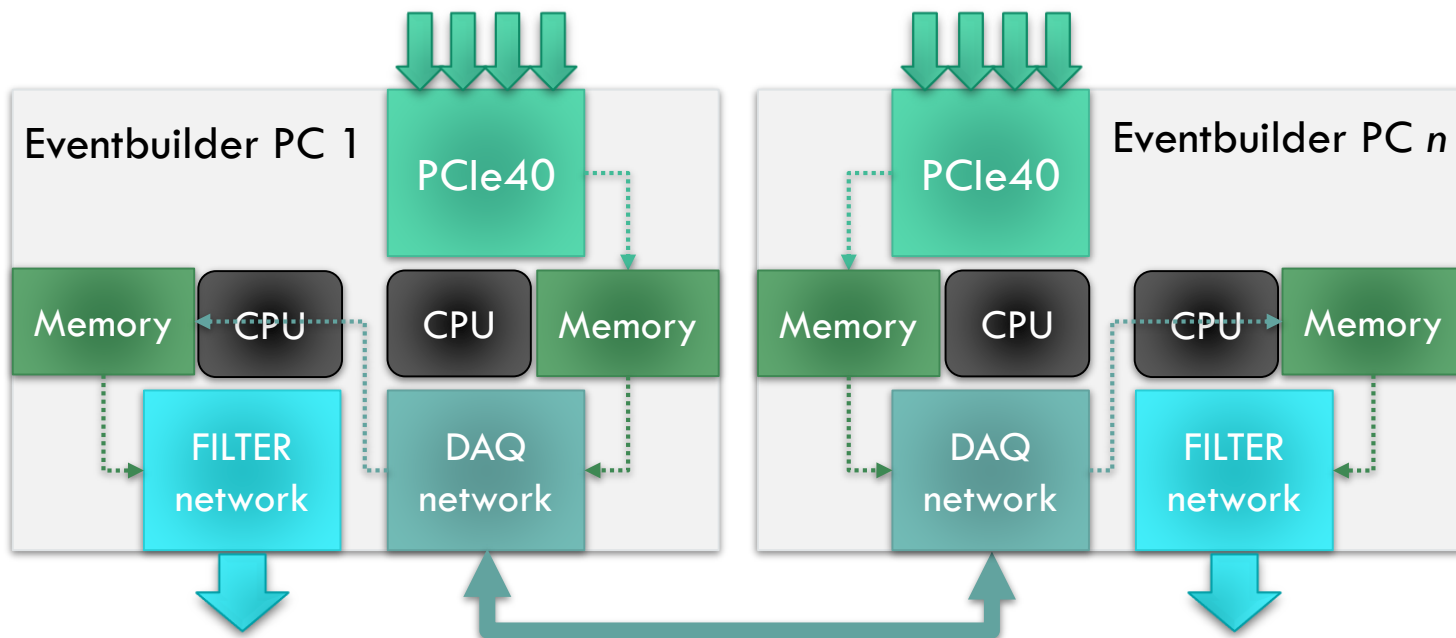
# Common PCIe40 firmware frame-work



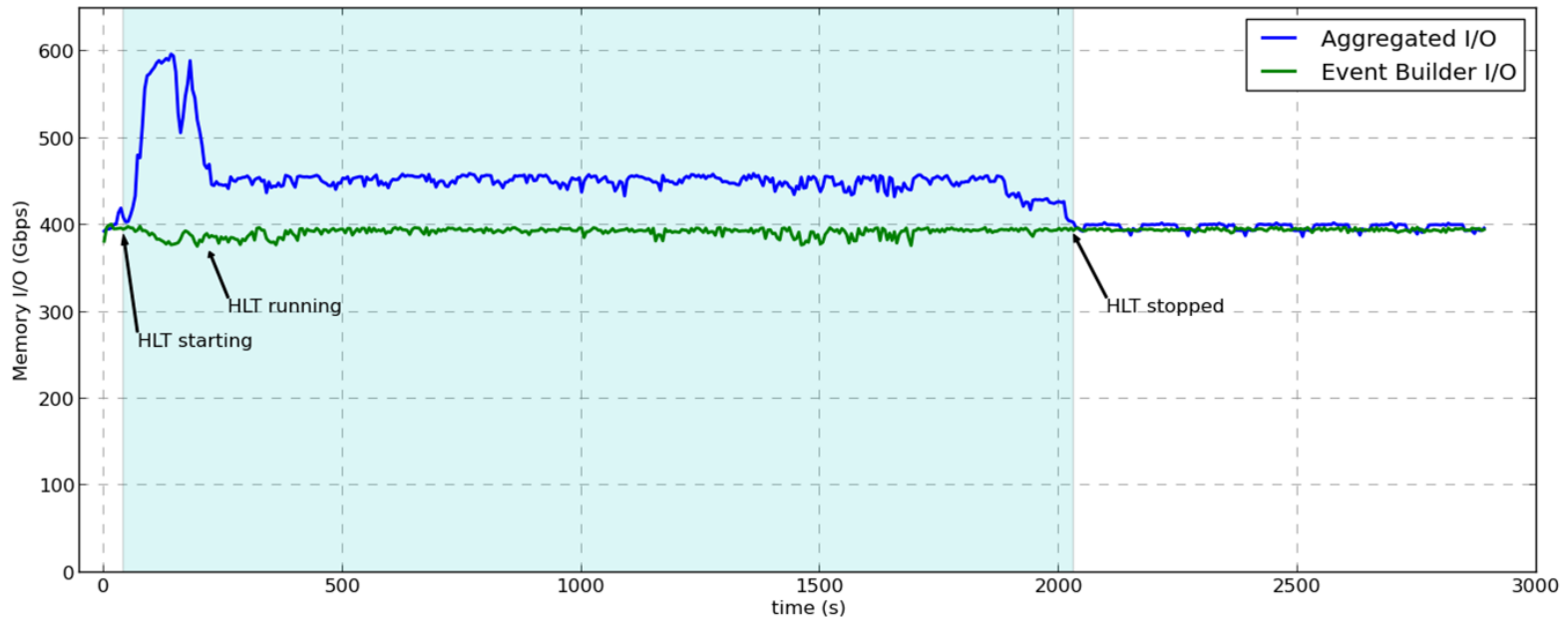


# Readout unit dataflow

- A single Readout unit must sustain  $\sim 400$  Gbps IO bandwidth
- Optimize memory bandwidth
  - Design for zero-copy operations and RDMA over the network
  - Organize dataflow according to topology and IO resources
  - Exploit full network bandwidth



# Folded event-builder network



Cons: Every node is readout and builder unit at the same time  $\rightarrow$  400 Gbit/s I/O

Limiting factor will be the memory bandwidth (not a problem)

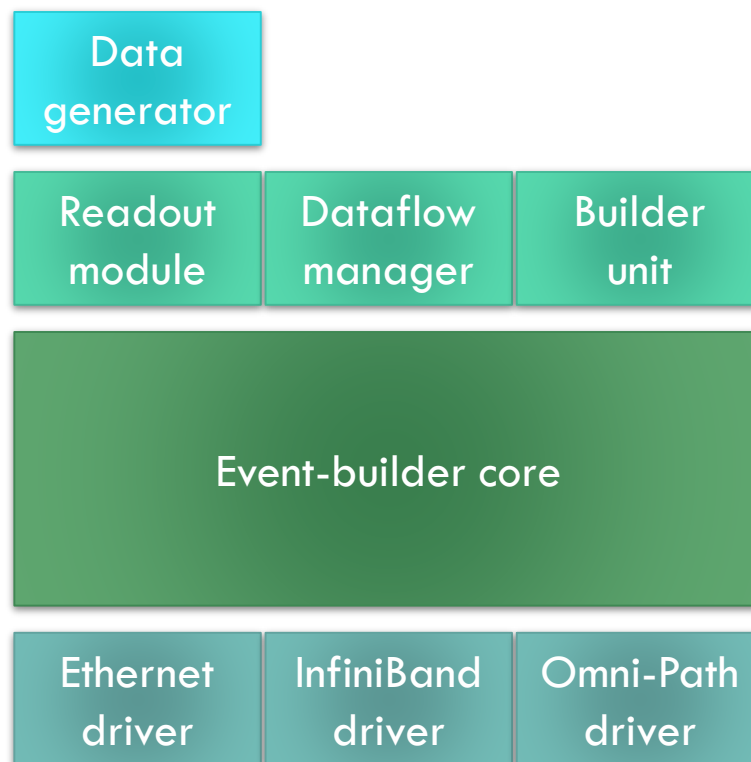
Pros: Fewer switch ports

Eventbuilder nodes see full events  $\rightarrow$  opportunistic usage of the CPU

Protocol conversion: eventbuilder and filter LAN can be different technology

# Event-building software (DAQPIPE)

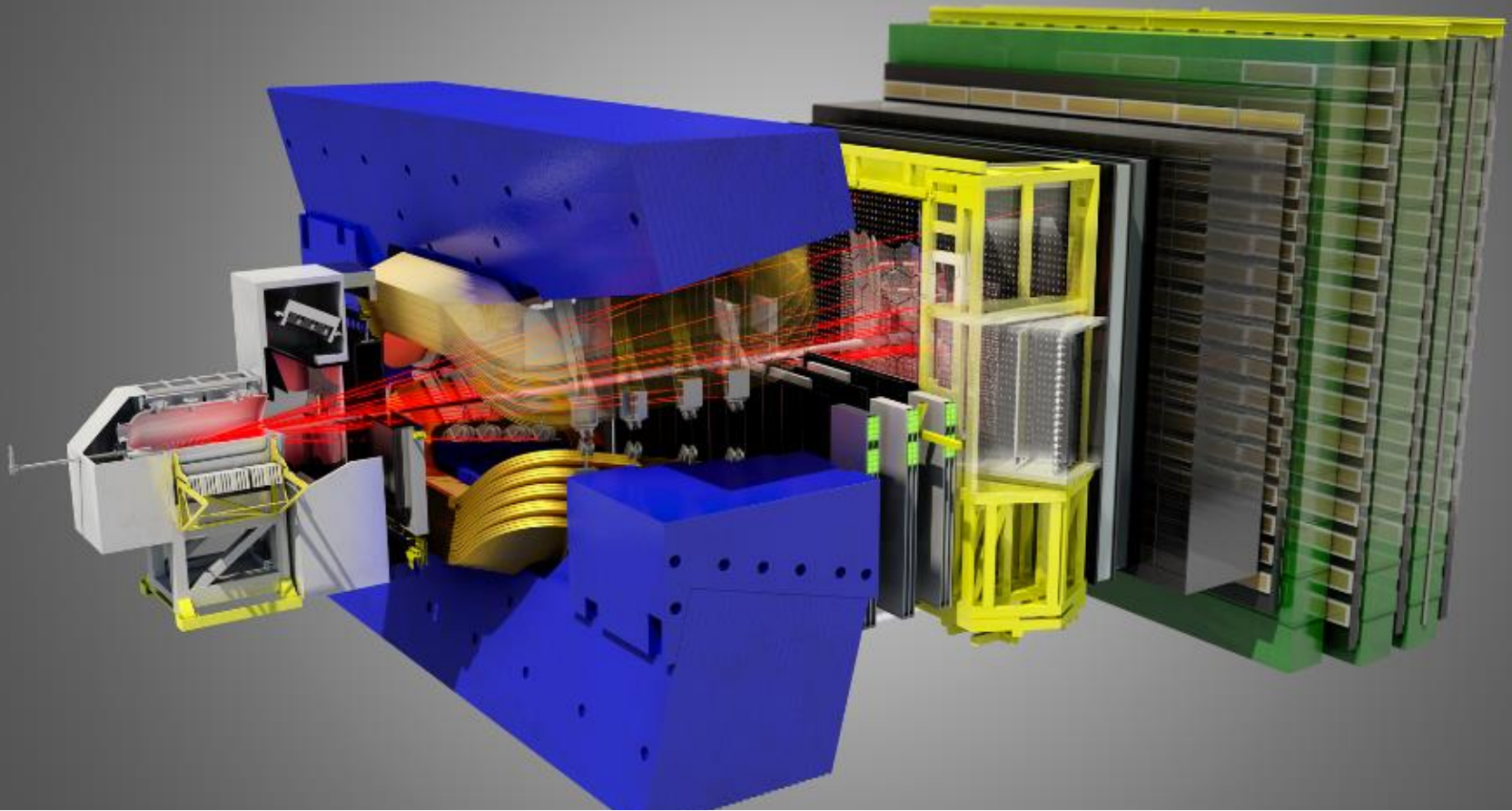
- Recreate distributed event-building dataflow of LHCb Run3
- Modular architecture, “drivers” for each network technology under evaluation
- Leverage existing HPC sites to assess scalability
- Close collaboration with the industry through CERN OpenLab
- Already achieving ~86 Gbps with Infiniband EDR and Intel OPA
  - Meets our target
  - Reduced scale setup
- External large-scale tests for InfiniBand EDR and Intel Omni-Path being prepared at HPC sites



# Future data centre at Point 8

	<b>Turnkey commercial solution</b> (Requires minimal support from CERN engineering groups)	<b>Leverage existing infrastructure at Point 8</b>
<b>Building</b>	Buy pre-fabricated containers from a commercial supplier	Accommodate the farm in an existing building (SX8 hall)
<b>Cooling</b>	Cooling solution depends on the vendor (e. g. free air cooling)	Passive rear-door heat exchangers using primary water from existing cooling towers <ul style="list-style-type: none"> <li>▪ Compatible with DCLC for hot spots</li> <li>▪ Test setup at the pit to evaluate performance with on site “warm” cooling water</li> </ul>

A review to decide the most cost-effective solution will take place in April.



**This will be the highest  
throughput data acquisition  
system ever built**

# Discussion slides

# Is it risking to become obsolete?

The DAQ depends only on:

1. Some kind of “server” capable of housing an FPGA based electronics board
2. A local area network technology

The DAQ is scalable horizontally and vertically

PCIe roadmap is compatibly safe until at least Run4 (Gen4 available probably from late 2017 onwards)

No dependence on exact server architecture, CPU architecture, GPGPUs, specific LAN technology



# Can we “unfold” the event-builder



In this case event-receiver and event-builder run on different PCs → go back to Run2 DAQ

Loose CPU power in event-builder PCs

But distribute the I/O better

Need more fully connected network ports

Again, cost will decide